FROM COGNITIVE-AFFECTIVE VULNERABILITIES TO DIGITAL AGENCY: A PEDAGOGICAL MODEL TO ADDRESS ALGORITHMIC MANIPULATION

Domenico Rubino[1]
Pegaso University
domenico.rubino@unipegaso.it

Anna Maria Mariani[2]
Pegaso University
annamaria.mariani@unipegaso.it

Double Blind Peer Review

Citation
Rubino, D., & Mariani, A. M. (2026). From cognitive-affective vulnerabilities to digital agency: A pedagogical model to address algorithmic manipulation. *Italian Journal of Health Education, Sport and Inclusive Didactics, 10*(3).

Doi: https://doi.org/10.32043/gsd.v10i3.1919

Copyright notice:
© 2023 this is an open access, peer-reviewed article published by Open Journal System and distributed under the terms of the Creative Commons Attribution 4.0 International, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

gsdjournal.it
ISSN: 3103-5744

ABSTRACT
This paper investigates the detrimental impact of social media algorithmic content distribution on users' cognitive and affective processes, while developing a pedagogical model to counter it. Evidence increasingly indicates that algorithms implemented by major platforms are designed to maximise engagement, intensifying emotional responses and systematically reinforcing reactive cognitive patterns. This phenomenon is commonly referred to as "algorithmic manipulation". Drawing on an extensive, up-to-date body of theoretical and empirical research, the study develops its analysis along three axes: the formulation of a systematic definition of algorithmic manipulation; the development of a classification of the domains of vulnerability to algorithmic systems; and a discussion of the limitations of normative and regulatory approaches. On this basis, it proposes an integrated pedagogical model aimed at strengthening critical, metacognitive, and emotional capacities to support agency.

KEYWORDS
Algorithmic manipulation; critical thinking; metacognition; emotional awareness; pedagogical model.

Received 2026-07-10
Accepted 2026-10-22
Published 2026-10-30

[1] Author contribution: Sections 1, 2, 4, and 6.
[2] Author contribution: Sections 3 and 5.

1. Introduction

When publicly questioned about the operation of algorithmic systems and the regulation of digital platforms, several leading figures in so-called Big Tech have adopted broadly overlapping rhetorical strategies. A recurring pattern in these exchanges is the tendency to recast demands for accountability as mere expressions of distrust toward innovation, thereby assimilating the contemporary debate on algorithmic manipulation and digital governmentality to the catastrophism surrounding the emergence of new technologies (Express News Service, 2026; Lui, 2017; U.S. Senate Committee on Commerce, Science, and Transportation, 2025). This analogy obscures the distinctive nature of digital algorithmic systems. A substantial body of empirical research shows that such computational architectures are not neutral tools, but adaptive structures deliberately designed to maximise user engagement, particularly by amplifying cognitive biases, reinforcing impulsive or reactive behavioural responses, and modulating emotion (Milli et al., 2025; Piccardi et al., 2025). Neither cinema nor the harnessing of electricity, for instance, made it possible to map individual cognitive vulnerabilities in real time, whereas contemporary social media platforms operate precisely through profiling and behavioural conditioning. A historical comparison between the two technological eras would amount to the logical fallacy of *false equivalence*. From the perspective of critical pedagogy, the use of such rhetoric to deflect criticism of the cognitive and emotional effects of digital algorithms constitutes a strategy for redefining the terms within which the issue is framed and problematised (Giroux, 1988). In the face of this growing erosion of accountability, further enabled by the broad liability protections afforded to digital platforms (17 U.S.C. § 512; 47

U.S.C. § 230), critical empowerment and agency become central concerns, particularly for younger generations, whose development is taking place within the opacity of the current digital ecosystem.

Educational interventions aimed at preventing algorithmic manipulation should seek to create the conditions for cognitive and affective awareness within the opaque and asymmetrical domain of human–platform interaction. The issue at stake is whether spaces for deliberate reflection can be preserved or created within environments engineered to steer attention and covertly amplify emotions. The central question is which tools can support practices of critical awareness that remain compatible with individuals' autonomy and personal values. Normative and disciplinary approaches (Bradford, 2024; Haidt, 2024) appear insufficient when treated as the sole response. Although they may perform a useful and sometimes necessary harm-containment function, they operate primarily according to a reactive, post-traumatic logic, and do not strengthen users' capacity to recognise and directly resist the cognitive-affective coercion exerted by algorithmic environments as it occurs. By constraining freedom of choice, such measures may trigger an oppositional response to perceived control (*reactance*) and thereby undermine adherence (Lukoff et al., 2022). The widely debated case of the UK Online Safety Act falls within this logic of intervention and highlights its potential risks, including the paradoxical institutionalisation of forms of content censorship and the erosion of privacy (ARTICLE 19, 2021).

Despite the breadth and significance of the existing research, the current debate on social media use (SMU) remains largely compartmentalised and tends to lack cohesion: theoretical discourse is often conceptually rich but offers few immediate educational implications (Susser et al., 2019); empirical studies address multiple dimensions of the phenomenon but remain

fragmented and lack a structured orientation, with metacognition, emotional awareness, and critical thinking treated as separate dimensions (Akbari et al., 2023; Ku et al., 2019; Piccerillo & Digennaro, 2025). Available research is dispersed across a heterogeneous range of disciplinary fields and lines of inquiry, including philosophy and democratic theory; privacy, surveillance, and autonomy; digital wellbeing; propaganda; psychology of compulsive use; media literacy. Algorithmic manipulation is not the prevailing scientific focus, while the popularity of the topic has encouraged the production of a considerable amount of grey literature, further increasing the fragmentation of the field. This imbalance also reflects the historical trajectory of the debate. The problem of algorithmic manipulation initially emerged within the social-media and political context of the crisis of the democratic information sphere, with particular attention being paid to phenomena such as misinformation and disinformation. The existence of a research gap concerning purposeful and non-coercive educational protocols is further borne out by the literature's limited adherence to international policy frameworks, which identify the development of interventions grounded in empowerment and agency as the principal way forward (Council of Europe, 2019; Organisation for Economic Co-operation and Development, 2021). Among the relatively few educational contributions, some address narrowly defined aspects of the issue, such as social media literacy limited to positivity bias, the stand-alone treatment of adaptive regulation of affective responses during social media use, and digital influence literacy (Schreurs & Vandenbosch, 2021, 2026; Senft, 2025). None of these contributions, however, approaches algorithmic manipulation as a unified pedagogical object or integrates critical, metacognitive, and emotional competencies within a framework designed to strengthen agency.

This paper develops an up-to-date and comprehensive conceptual model of pedagogical intervention centred on the formative conditions required for the exercise of agency in digital environments. Through a comparative mapping of the scientific literature on algorithmic manipulation and social media use (critical algorithm studies, algorithmic literacy, media education, digital wellbeing, platform governance) it examines the impact of algorithmic mechanisms of content selection and distribution on social media users' cognitive and affective processes. The analytical framework underpinning the proposal is organised around three axes: first, the formulation of a systematic and analytical definition of algorithmic manipulation through the identification of its distinctive properties and main operating mechanisms; second, the development of a classification of the domains of vulnerability to algorithmic systems (critical-cognitive, emotional-regulatory, metacognitive-regulatory); and third, a discussion of the limitations of normative and regulatory approaches, including through the examination of specific measures and events, with a focus on the effects of the implementation of the UK Online Safety Act. On this basis, the model identifies critical, metacognitive, and emotional empowerment as an effective practical pathway for strengthening reflective autonomy and emotional awareness in algorithmically mediated digital contexts.

2. Algorithmic Manipulation: Definition and Areas of Risk

Algorithmic manipulation is the intentional use of adaptive algorithmic systems, including those deployed on social media platforms and in AI-based systems, to covertly influence users' decision-making processes (Faraoni, 2023; Susser et al., 2019). It involves more than exposure to false or persuasive content; rather, it constitutes a deeper, structural modulation of digital

experience. By systematically collecting interaction data and behavioural signals, algorithms build predictive models of users' responses (Yeung, 2017) and use them to dynamically maximise engagement with the platform (Milli et al., 2025). Users are thereby exposed to personalised input–output dynamics that pervasively shape their cognitive and affective dispositions, potentially to the point of constraining their evaluative processes. Because the decision-making mechanisms linking user-generated input (clicks, scrolling, posts) to algorithmically generated output (content selection, recommendations, personalised advertising) remain opaque, users do not clearly perceive the resulting restriction of their decision-making autonomy (Felzmann et al., 2019).

One of the central challenges in defining algorithmic manipulation is distinguishing it from interpersonal forms of manipulation. Interpersonal manipulation depends on the direct intervention of the manipulator and is therefore limited by the reach of individual action and by a comparatively weaker capacity to adapt to the recipient's responses. Algorithmic manipulation, by contrast, is characterised by high-speed autonomous operation (*automation*), simultaneous large-scale functioning (*scalability*), continuous adaptive modulation (*feedback loops*) (Yeung, 2017), and opaque decision-making processes (*black box*) (Wachter et al., 2018). In particular, the parallelisation of targeting operations gives manipulative action an unprecedented reach. A potentially global user base can be exposed, at speeds unattainable through interpersonal manipulation, to personalised sequences of interaction designed to maximise time spent on the platform and, more broadly, to achieve its operational objectives. The system continuously learns which stimuli retain attention, while the underlying decision-making process remains opaque. Users encounter the outputs of algorithmic selection without meaningful access to the criteria governing it. The literature therefore

distinguishes algorithmic from interpersonal manipulation on the grounds that it relocates manipulative capacity from a bounded interpersonal relationship to an automated infrastructure that continuously modulates individual responses.

The structured mechanisms most commonly associated with algorithmic manipulation include the dynamic and personalised steering of behaviour through digital choice architectures (*hypernudging*) (Yeung, 2017); informational isolation produced by algorithmic personalisation (*filter bubbles*) (Pariser, 2011); and closed environments in which beliefs are mutually reinforced through systems that recommend both content and connections between users (*echo chambers*) (Sunstein, 2001). In early sociopolitical and media debates, the most immediate risks associated with these dynamics were identified as a decline in informational pluralism and interference with democracy and elections. Later, the scope of inquiry broadened to systematically include users' cognitive, affective, and behavioural processes. Neuropsychological approaches have drawn on constructs such as *algorithmic dependency* and *compulsive engagement*, usually within broad, predominantly reactive taxonomic frameworks (Aamir et al., 2026; Shannon et al., 2022). At the same time, empirical research has also begun to break down the cognitive-affective effects of algorithmic manipulation into more discrete analytically defined components. Studies on the cognitive and behavioural correlates of social media use identify specific self-regulatory impairments among users (automaticity, loss of intentionality) (Du et al., 2019), as well as recognisable forms of reactive/impulsive engagement (reactive consumption, immediate reward seeking, impulsive clicking/scrolling) (Turel & Qahri-Saremi, 2018) or compulsive engagement (compulsive checking, inability to disengage) (Tran et al., 2019). With regard to emotional effects, analyses of engagement-based

ranking systems and experiments conducted on real-world feeds show that the algorithmic amplification of highly affect-laden content or content marked by partisan animosity increases users' negative emotions and affective polarisation (Milli et al., 2025; Piccardi et al., 2025).

Within this framework, three areas of vulnerability to algorithmic systems can be identified. To begin with, the cognitive and behavioural effects associated with impaired self-regulation fall into two broad areas. The first is the critical-cognitive domain and concerns compromised evaluative processes and the difficulty of recognising algorithmic mediation, reconstructing its criteria and assessing its purposes. The second is the metacognitive-regulatory dimension and includes automaticity, loss of intentionality, compulsive engagement, difficulty disengaging, and reactive or impulsive modes of interaction. Alongside these lies the emotional-regulatory area, which is linked to the amplification and intensification of affective responses. These three domains can therefore be formalised as vulnerabilities in critical thinking, metacognition, and emotional awareness. The literature has not yet organised these phenomena within a systematic pedagogical framework grounded in a non-reactive epistemological approach. Psychological and behavioural research tends to treat self-regulatory domains as *boundary conditions*, emphasising their diagnostic significance while leaving their transformative potential underdeveloped (Reinecke et al., 2022). This reveals an epistemic continuity between the way algorithmic manipulation is framed in the literature and the strategies through which disciplinary approaches seek to counter its social effects. Both rest on a normative, post-traumatic conception of intervention.

3. Functions and Limitations of Normative Approaches

The normative and disciplinary interventions designed to counter the dysfunctional effects of social media and algorithmic manipulation emerged alongside the early sociopolitical debate on the crisis of the democratic information sphere. Within this context, platform power was addressed chiefly as a matter of public order and social regulation, with the emphasis placed on institutional oversight and external control. In practice, this response has generally taken the form of government action rather than structured pedagogical intervention. Its function is to manage access and exposure to social media platforms from the outside by imposing a coercive regulatory system that may include age thresholds, verification requirements, filters, bans, moderation procedures, and compliance mechanisms. Normative approaches do not speak the language of agency and empowerment promoted by the relevant international policy frameworks (Council of Europe, 2019; Organisation for Economic Co-operation and Development, 2021), nor do they advance a research agenda centred on the proactive development of critical thinking, metacognition, and emotional awareness. Instead, they translate the problems of algorithmic manipulation and the adverse effects of social media into an administrative grammar of containment, externally setting the boundaries within which users may interact with platforms, in a reactive and pedagogically ineffective manner.

From a pedagogical perspective, these interventions are structurally limited. They operate primarily as reactive and remedial mechanisms, intervening only after harm has been formally identified as requiring regulation, and merely seeking to reduce further exposure to potentially detrimental environments. Their logic is therefore predominantly heteroregulatory: protection is achieved through external restriction instead of by developing users' own capacities for

recognition, self-regulation, critical thinking, and emotional awareness. Research on digital self-control tools points to both limited effectiveness and problematic acceptability. Among the shortcomings identified, externally imposed restrictions may provoke psychological reactance — a drive to restore a freedom perceived as having been taken away — and thereby undermine the very regulatory objective they are intended to achieve (Lukoff et al., 2022).

The limitations of normative approaches are equally apparent at the social level. The United Kingdom's Online Safety Act provides an emblematic and widely debated case of paradoxical effects. The Act became law on 26 October 2023 (Online Safety Act, 2023). It assigns Ofcom oversight of social and video platforms, forums, messaging services, search engines, and other spaces in which users publish content accessible from the United Kingdom, imposing risk assessments, moderation systems, the removal of illegal content, and age-assurance measures. The Act targets effects of algorithmic systems that are classified as harmful. Ofcom may require changes to services and exact compliance through financial penalties or court-ordered services blocking. The framework also incorporates criminal enforcement measures against users, up to and including arrest, within an externally directed and coercive model of intervention. The effectiveness claimed for the Online Safety Act rests on three controversial presuppositions that have been extensively debated in the literature: onerous procedural duties and selective automated moderation, including user profiling, can translate directly into greater safety without simultaneously producing surveillance and censorship (ARTICLE 19, 2021); a univocal definition of *hate speech* can be formally codified and legally imposed without paradoxically violating freedom of expression (Brown & Sinclair, 2023); the resulting system can be implemented without infringing users' privacy (Open Rights Group, 2023). Reports of restrictive measures affecting even very

young people (North Wales Police, 2026) illustrate a reversal of the Act's stated purpose: the minor designated as the object of protection becomes the target of repression. Such a logic is incompatible with a pedagogical perspective, which seeks to foster inclusion rather than create new forms of criminal liability.

4. From Vulnerability to Agency: A Pedagogical Model to Counter Algorithmic Manipulation

Instead of relying on heteroregulatory and post-traumatic containment, a pedagogical model centred on critical empowerment requires the vulnerabilities framed as sociopolitical and diagnostic by the literature to be recast as educational objectives. To this extent, the present model reorganises the main domains of algorithmic exposure within a proactive pedagogical framework designed to counter platform-mediated manipulation. Cognitive-emotional effects such as compromised evaluative processes, loss of intentionality, automatic interaction, impulsivity, and emotional dysregulation are not merely targets of intervention; viewed from another angle, they identify the agentic capacities that educational action must strengthen. Impaired evaluation and difficulty recognising algorithmic mediation call for a stronger critical understanding of how algorithmic systems operate. Automaticity, loss of intentionality, compulsive engagement, and reactive or impulsive modes of interaction call for metacognitive practices of observation and self-regulation. The amplification of affective responses calls for the development of emotional awareness. This model brings these three dimensions together while preserving their distinct functions and grounding them in users' concrete experience of algorithmically mediated digital environments.

The pedagogical architecture developed in this study comprises three interconnected domains. Each derives from the theoretical and operational systematisation of a distinct body of empirical findings, considered both for the vulnerability it documents and the space for intervention it leaves open. To begin with, research on algorithmic literacy shows that opaque decision-making processes prevent users from reconstructing the criteria that organise the platform ecosystem and the purposes embedded in platforms (Rader & Gray, 2015). Psychometric studies developing and validating algorithmic literacy and awareness scales make the mediation structuring the digital environment recognisable, while remaining largely focused on understanding the system, and do not always translate their findings into a practical capacity to navigate the interaction (Dogruel et al., 2022; Zarouali et al., 2021). The model identifies this gap as its first educational problem. It corresponds, as its negative formulation, to the critical-cognitive domain, which concerns the ability to recognise algorithmic mediation and understand its main modes of operation. Users must be able to distinguish the content displayed from the criteria that determined its selection by scrutinising the role of profiling, filtering, ranking, recommendation, and personalisation. Educational interventions based on narrative scripts and algorithm-awareness modules have increased both algorithmic awareness and critical disposition, strengthening participants' ability to question system reliability and assess the degree of control they retain over their personal data (Markus et al., 2025; Theophilou et al., 2023). Such findings confirm both the educational relevance of the critical-cognitive domain and the possibility of strengthening the corresponding abilities through educational intervention. The model assigns this level the function of addressing the vulnerability produced by algorithmic opacity through the development of three critical-cognitive resources: detecting the presence of algorithmic mediation as it operates, identifying the

criteria that determine content selection, and evaluating the economic and behavioural purposes they are designed to serve. These capacities are a necessary condition for transforming unreflective exposure to platform outputs into an understanding that enables users to orient themselves within the interaction.

Next, the metacognitive-regulatory domain concerns monitoring one's own patterns of interaction: initial intention, distribution of attention, time spent on the platform, recurrence of automatic actions, and difficulty disengaging. Empirical studies show that self-regulatory competence is impaired in social media use, particularly by habitual platform checking, perceived ubiquitous connectivity, as well as notification-induced disruption (Du et al., 2019). Among adolescents, automaticity in social media use is also associated with smartphone-checking frequency and, in different individual configurations, with goal-directed activity postponement (Meier et al., 2023). These studies identify the dynamics and consequences of impaired self-regulation but remain largely descriptive, without translating the assessed indicators into practices enabling users to recognise them during platform engagement. The model addresses this gap as its second educational problem, which finds its positive counterpart in the metacognitive-regulatory domain. It involves the ability to self-monitor: initial intention, shifts in attention, duration of use, recurring automatisms, and difficulty disengaging. This function is supported by interventions that have used advanced goal-setting and auto-observation to increase perceived self-control and autonomy, reducing absentminded social media use (Davis et al., 2023), as well as distraction in smartphone and social network use (Throuvala et al., 2020). The model therefore positions the development of the metacognitive-regulatory domain as a response to the vulnerabilities in self-regulation and counters loss of control by fostering three

specific capacities: recognising the purpose of accessing the platform, monitoring discrepancies between intention and behaviour, identifying and interrupting automatic or compulsive patterns that prevent users disengaging from the platform. The function of this level of analysis is not to impose an external reduction in time spent online, but to enable users to observe the process through which interaction progressively drifts away from their initial aims.

Finally, the emotional-regulatory domain concerns recognising the emotional states activated or intensified during digital experience. Experimental studies of algorithmic content modulation show that altering the affective valence of posts can shape users' subsequent emotional expression, and that reranking a feed can cause changes in affective polarisation and negative emotions (Kramer et al., 2014; Piccardi et al., 2025). These studies document emotional responses; however, they do not develop practices supporting users in recognising and evaluating affective activation as it occurs during interaction with the platform. The model treats this gap as its third educational problem and assigns the emotional-regulatory level the task of making observable both changes in the intensity of affective states and the relationship between those changes and immediate responses. Evidence that these capacities can be modified comes from affective social media literacy interventions and systematic reviews, which respectively report reductions in negative emotions and preliminary benefits for affect regulation. They identify adaptive emotional awareness as a key competence that should be strengthened (Schreurs & Vandenbosch, 2026; Thomson et al., 2025). The model therefore situates the development of the emotional-regulatory domain as a strategy for addressing vulnerabilities associated with emotional dysregulation and counteracts them by developing the following set of competences: identifying and differentiating

affective states, recognising their intensification during platform use, suspending automatic responses to assess whether they are consistent with personal intentions and goals. Emotional awareness thus plays a constitutive role in enabling a non-externally directed understanding of the processes through which the platform steers action.

The three domains are operationally integrated through a four-stage sequence: recognising algorithmic mediation; monitoring cognitive, behavioural, and emotional responses; reflectively evaluating the interaction; redirecting action (Fig. 1). The first stage consists in identifying the layer of algorithmic selection between user activity and the digital environment presented in response, thereby disrupting the perception of the platform as a neutral or spontaneously organised space. The second introduces metacognitive observation of the experience as it unfolds: users notice divergences between their initial purpose in accessing the platform and their emerging behaviour, recognise possible automatisms, and identify changes in their emotional state. The third stage relates these responses to personal values and consciously pursued aims. The key pedagogical question extends beyond the truthfulness of content or the amount of time spent online and asks whether the interaction is consistent with what users themselves regard as meaningful. The fourth stage concerns the possibility of modifying or interrupting the action, thereby recovering a margin of choice in relation to the sequence of stimuli generated by the platform.

The model thus complements the existing literature by supplying a systematic pedagogical response to the need emerging from it. Its architecture, categories, and the logical relations among its elements constitute an original framework for guiding educational analysis and interventions against platform-mediated manipulation, with broader application to the adverse cognitive-emotional repercussions of social media use. While existing studies provide valuable

empirical evidence on specific effects and patterns of interaction, the present research builds on this evidence to develop a structured operational model that advances such lines of inquiry within a coherent pedagogical perspective. As with any conceptual model, the framework proposed here requires further empirical application and validation in concrete educational settings.

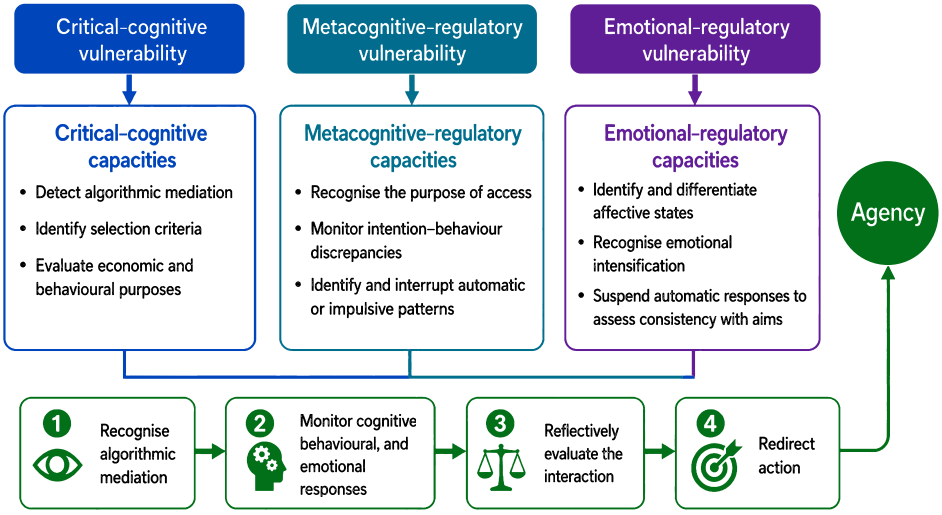


Fig. 1. Structure of the pedagogical model: from vulnerability to agency.

5. Discussion

The main contribution of this paper is the development of a pedagogical framework that recasts as educational objectives the cognitive-emotional vulnerabilities documented in the scientific literature on algorithmic manipulation and social media use. Thus, it guides educational *praxis* in countering the detrimental effects of algorithmically mediated digital environments. The proposal is grounded in a systematic and comprehensive definition of the phenomenon as the intentional modulation of users' behaviour and cognitive-emotional experience within algorithmically mediated

digital environments, specifically through automation, scalability, continuous adaptive modulation, and opaque decision-making processes. The literature has examined its effects largely within separate fields, treating algorithmic opacity, automatic use, loss of intentionality, compulsive engagement, and emotional intensification as discrete issues, without producing an operational model that consistently integrates them within an agency-centred pedagogy. This proposal addresses the gap through the strengthening of three core dimensions: critical thinking, metacognition, and emotional awareness. Rather than treating users' vulnerabilities simply as diagnostic indicators or as conditions to be contained externally, it reframes them as educational problems that call for the fostering of specific capacities.

The comparison with normative and disciplinary approaches reveals a structural contrast between their logic of intervention and that of the proposed framework. Coercive regulatory measures impose external control while leaving untouched the internal processes that make individuals vulnerable to manipulation. They operate reactively and post hoc: intervention occurs once harm has been formally recognised as warranting regulation, remaining primarily confined to reducing further exposure to potentially harmful environments, and thereby translating an educational problem into an administrative grammar of control. Such approaches neither develop critical thinking or cognitive-affective awareness nor make automatisms recognisable. They do not strengthen the ability to distinguish between initial intention and the behaviour produced as interaction unfolds. Far from being marginal anomalies, the reactance elicited by restrictions and the repressive effects associated with the Online Safety Act are coherent manifestations of a paradigm that replaces the development of autonomy with coercion. The paradox whereby the minor formally designated as the recipient of protection

becomes an object of control and criminal liability demonstrates the incompatibility of this framework with a genuinely pedagogical perspective.

As a pedagogical alternative, the proposed model addresses algorithmic manipulation in the very domain upon which it acts: the formation and modulation of individual responses. Critical thinking, metacognition, and emotional awareness are organised into a sequence that moves from recognising the algorithmic conditions of the digital environment to monitoring one's own responses and consciously redirecting action. The standard of success is not compliance with heteroregulatory restrictions but empowerment through ability development for understanding and evaluating processes of cognitive-affective constraint that displace behaviour from its initial purposes. The model's pedagogical effectiveness lies in its anticipatory and transformative character: it does not wait for harm to occur and then seek to contain its effects; it fosters, in advance and through practice, the capacities required to identify and disrupt cognitive-affective constraint.

6. Conclusions

Responses to algorithmic manipulation should not be reduced to the normative imposition of externally directed measures, such as access controls and content moderation. These disciplinary interventions may perform a containment function, but they do not strengthen the capacities required to counter the cognitive-affective effects generated by digital platforms. The model seeks to shift the focus of intervention from the external control of digital experience to the development of users' capacity to navigate it consciously. The domains over which manipulation exerts its strongest hold — critical understanding, metacognitive regulation, and emotional awareness — thus also become the primary sites of pedagogical transformation. Their operational integration

creates a space for reflection between algorithmically generated stimuli and individual response, within which users can recognise and assess the interaction, and redirect it in accordance with consciously held values and personal aims. Countering algorithmic manipulation thereby assumes a properly pedagogical function: strengthening the cognitive and emotional conditions that support the exercise of agency.

References

Aamir, A., Munir, S., Hafsa, H., & Khalid, W. (2026). Neural parasitism: Could adaptive artificial intelligence systems incrementally reconfigure human neural plasticity and challenge the foundations of cognitive autonomy? *Annals of Medicine & Surgery, 88*(3), 2240–2244. https://doi.org/10.1097/MS9.0000000000004677

Akbari, M., Hossein Bahadori, M., Khanbabaei, S., Boruki Milan, B., Horvath, Z., Griffiths, M. D., & Demetrovics, Z. (2023). Metacognitions as a predictor of problematic social media use and internet gaming disorder: Development and psychometric properties of the Metacognitions about Social Media Use Scale (MSMUS). *Addictive Behaviors, 137*, Article 107541. https://doi.org/10.1016/j.addbeh.2022.107541

ARTICLE 19. (2021, July 26). *UK: Draft Online Safety Bill poses serious risk to free expression*. https://www.article19.org/resources/uk-draft-online-safety-bill-poses-serious-risk-to-free-expression/

Bradford, A. (2024). The false choice between digital regulation and innovation. *Northwestern University Law Review, 119*(2), 377–454. https://scholarlycommons.law.northwestern.edu/nulr/vol119/iss2/3/

Brown, A., & Sinclair, A. (2023). *Hate speech frontiers: Exploring the limits of the ordinary and legal concepts*. Cambridge University Press. https://doi.org/10.1017/9781009357111

Communications Decency Act of 1996, 47 U.S.C. § 230 (1996). https://www.govinfo.gov/link/uscode/47/230

Council of Europe. (2019). *CM/Rec(2019)10 – Recommendation of the Committee of Ministers to member States on developing and promoting digital citizenship education*.

https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=090000168098de08

Davis, K., Slovak, P., Landesman, R., Pitt, C., Ghajar, A., Schleider, J. L., Kawas, S., Portillo, A. G. P., & Kuhn, N. S. (2023). Supporting teens' intentional social media use through interaction design. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (pp. 322–334). Association for Computing Machinery. https://doi.org/10.1145/3585088.3589387

Dogruel, L., Masur, P., & Joeckel, S. (2022). Development and validation of an algorithm literacy scale for internet users. *Communication Methods and Measures, 16*(2), 115–133. https://doi.org/10.1080/19312458.2021.1968361

Du, J., Kerkhof, P., & van Koningsbruggen, G. M. (2019). Predictors of social media self-control failure: Immediate gratifications, habitual checking, ubiquity, and notifications. *Cyberpsychology, Behavior, and Social Networking, 22*(7), 477–485. https://doi.org/10.1089/cyber.2018.0730

Express News Service. (2026, March 7). Sam Altman at Adda: "I don't think there should be any single superintelligence in the world". *The Indian Express*. https://indianexpress.com/article/india/sam-altman-at-adda-i-dont-think-there-should-be-any-single-superintelligence-in-the-world-10568974/

Faraoni, S. (2023). Persuasive technology and computational manipulation: Hypernudging out of mental self-determination. *Frontiers in Artificial Intelligence, 6*, Article 1216340. https://doi.org/10.3389/frai.2023.1216340

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1), 1–14. https://doi.org/10.1177/2053951719860542

Giroux, H. A. (1988). *Teachers as intellectuals: Toward a critical pedagogy of learning*. Bergin & Garvey.

Haidt, J. (2024). *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness*. Penguin Press.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Ku, K. Y. L., Kong, Q., Song, Y., Deng, L., Kang, Y., & Hu, A. (2019). What predicts adolescents' critical thinking about real-life news? The roles of social media news consumption and news media literacy. *Thinking Skills and Creativity, 33*, Article 100570. https://doi.org/10.1016/j.tsc.2019.05.004

Lui, K. (2017, July 25). Mark Zuckerberg says Elon Musk is drumming up "doomsday scenarios" about artificial intelligence. *Fortune*. https://fortune.com/2017/07/25/mark-zuckerberg-elon-musk-artificial-intelligence/

Lukoff, K., Lyngs, U., & Alberts, L. (2022). Designing to support autonomy and reduce psychological reactance in digital self-control tools. CHI '22 Workshop. https://ulriklyngs.com/pdfs/lukoff-lyngs-alberts-2022-dscts_for_autonomy.pdf

Markus, A., Baumann, M., Pfister, J., Hotho, A., Carolus, A., & Wienrich, C. (2025). The impact of algorithm awareness training on competent interaction with intelligent voice assistants. *Discover Education, 4*, Article 125. https://doi.org/10.1007/s44217-025-00522-6

Meier, A., Beyens, I., Siebers, T., Pouwels, J. L., & Valkenburg, P. M. (2023). Habitual social media and smartphone use are linked to task delay for some, but not all,

adolescents. *Journal of Computer-Mediated Communication, 28*(3), Article zmad008. https://doi.org/10.1093/jcmc/zmad008

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus, 4*(3), pgaf062. https://doi.org/10.1093/pnasnexus/pgaf062

North Wales Police. (2026, April 16). *Teenager arrested on suspicion of online offences*. https://www.northwales.police.uk/news/north-wales/news/news/2026/april/teenager-arrested-on-suspicion-of-online-offences2/

Online Copyright Infringement Liability Limitation Act, 17 U.S.C. § 512 (1998). https://www.govinfo.gov/link/uscode/17/512

Online Safety Act 2023, c. 50 (UK). https://www.legislation.gov.uk/ukpga/2023/50

Open Rights Group. (2023, September 5). *Online Safety Bill: US and UK campaigners warn of dangers of age verification*. https://www.openrightsgroup.org/press-releases/online-safety-bill-us-and-uk-campaigners-warn-of-dangers-of-age-verification/

Organisation for Economic Co-operation and Development. (2021). *Recommendation of the Council on children in the digital environment* (OECD/LEGAL/0389). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0389

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. The Penguin Press.

Piccardi, T., Saveski, M., Jia, C., Hancock, J. T., Tsai, J. L., & Bernstein, M. S. (2025). Reranking partisan animosity in algorithmic social media feeds alters affective

polarization. *Science, 390*(6776), eadu5584. https://doi.org/10.1126/science.adu5584

Piccerillo, L., & Digennaro, S. (2025). Adolescent social media use and emotional intelligence: A systematic review. *Adolescent Research Review, 10*(2), 201–218. https://doi.org/10.1007/s40894-024-00245-z

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 173–182). Association for Computing Machinery. https://doi.org/10.1145/2702123.2702174

Reinecke, L., Gilbert, A., & Eden, A. (2022). Self-regulation as a key boundary condition in the relationship between social media use and well-being. *Current Opinion in Psychology, 45*, Article 101296. https://doi.org/10.1016/j.copsyc.2021.12.008

Schreurs, L., & Vandenbosch, L. (2021). Introducing the social media literacy (SMILE) model with the case of the positivity bias on social media. *Journal of Children and Media, 15*(3), 320–337. https://doi.org/10.1080/17482798.2020.1809481

Schreurs, L., & Vandenbosch, L. (2026). Enhancing adolescents' emotion regulation in the social media context: A cluster randomized controlled trial of the Vibe Check social media literacy intervention. *Educational Psychology, 46*(5), 726–744. https://doi.org/10.1080/01443410.2025.2593545

Senft, T. M. (2025). Who cares how information feels? A call for digital influence literacy. *Asian Bioethics Review, 17*(3), 477–493. https://doi.org/10.1007/s41649-024-00350-0

Shannon, H., Bush, K., Villeneuve, P. J., Hellemans, K. G. C., & Guimond, S. (2022). Problematic social media use in adolescents and young adults: Systematic review

and meta-analysis. *JMIR Mental Health, 9*(4), e33450. https://doi.org/10.2196/33450

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review, 8*(2). https://doi.org/10.14763/2019.2.1410

Theophilou, E., Lomonaco, F., Donabauer, G., Ognibene, D., Sánchez-Reina, R. J., & Hernández-Leo, D. (2023). AI and narrative scripts to educate adolescents about social media algorithms: Insights about AI overdependence, trust and awareness. In O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, & T. Papathoma (Eds.), *Responsive and sustainable educational futures* (pp. 415–429). Springer. https://doi.org/10.1007/978-3-031-42682-7_28

Thomson, A., Lawrence, E., Sharxhi, E., Oliver, B. R., Wright, B., & Hosang, G. (2025). Self-directed digital interventions for the improvement of emotion regulation–acceptability and feasibility for adolescents: Systematic review. *BJPsych Open, 11*(6), Article e270. https://doi.org/10.1192/bjo.2025.10888

Throuvala, M. A., Griffiths, M. D., Rennoldson, M., & Kuss, D. J. (2020). Mind over matter: Testing the efficacy of an online randomized controlled trial to reduce distraction from smartphone use. *International Journal of Environmental Research and Public Health, 17*(13), Article 4842. https://doi.org/10.3390/ijerph17134842

Tran, J. A., Yang, K. S., Davis, K., & Hiniker, A. (2019). Modeling the engagement-disengagement cycle of compulsive phone use. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 312, pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300542

Turel, O., & Qahri-Saremi, H. (2018). Explaining unplanned online media behaviors: Dual system theory models of impulsive use and swearing on social networking

sites. *New Media & Society, 20*(8), 3050–3067. https://doi.org/10.1177/1461444817740755

U.S. Senate Committee on Commerce, Science, and Transportation. (2025, May 8). *Winning the AI race: Strengthening U.S. capabilities in computing and innovation* [Hearing]. https://www.commerce.senate.gov/meetings/winning-the-ai-race-strengthening-u-s-capabilities-in-computing-and-innovation/

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society, 20*(1), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

Zarouali, B., Boerman, S. C., & de Vreese, C. H. (2021). Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale). *Telematics and Informatics, 62*, Article 101607. https://doi.org/10.1016/j.tele.2021.101607